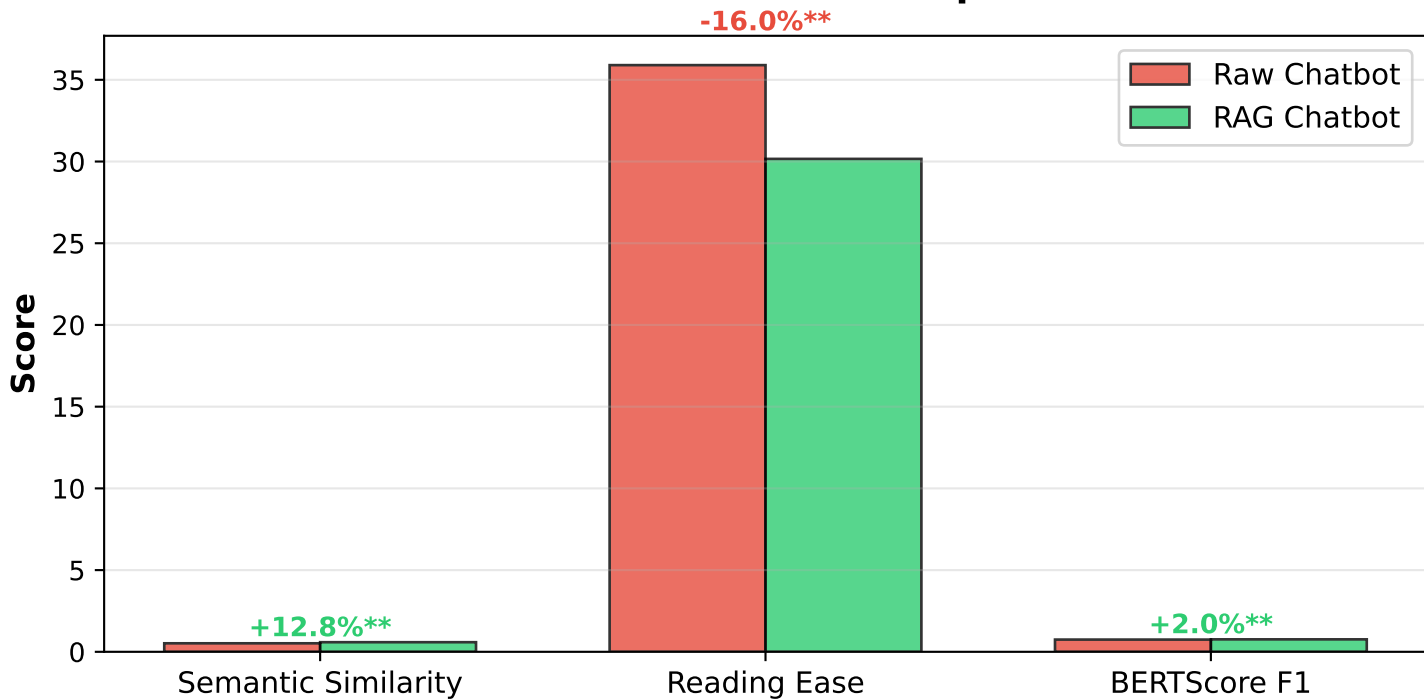


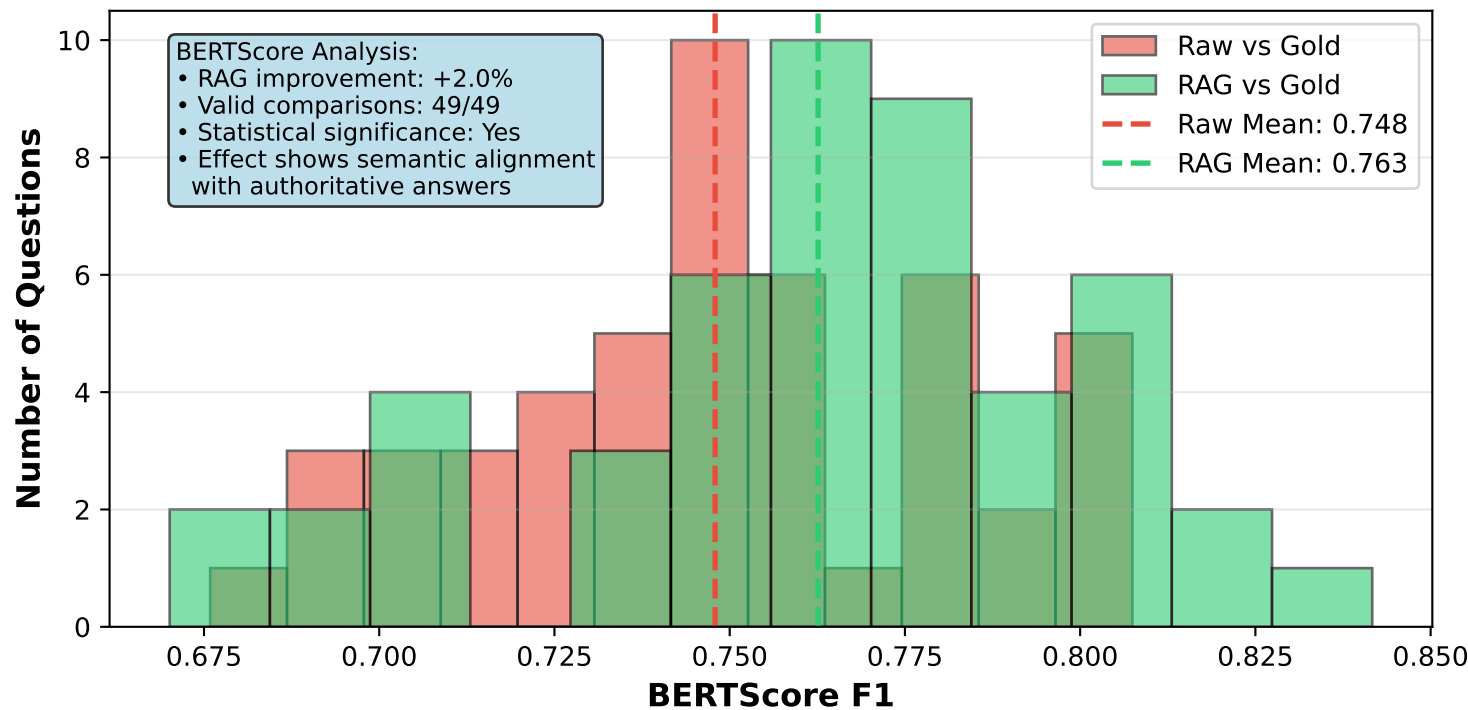
CGT Chatbot Comprehensive Analysis: RAG vs Raw Performance

BERTScore vs Gold Standard + Category Breakdown

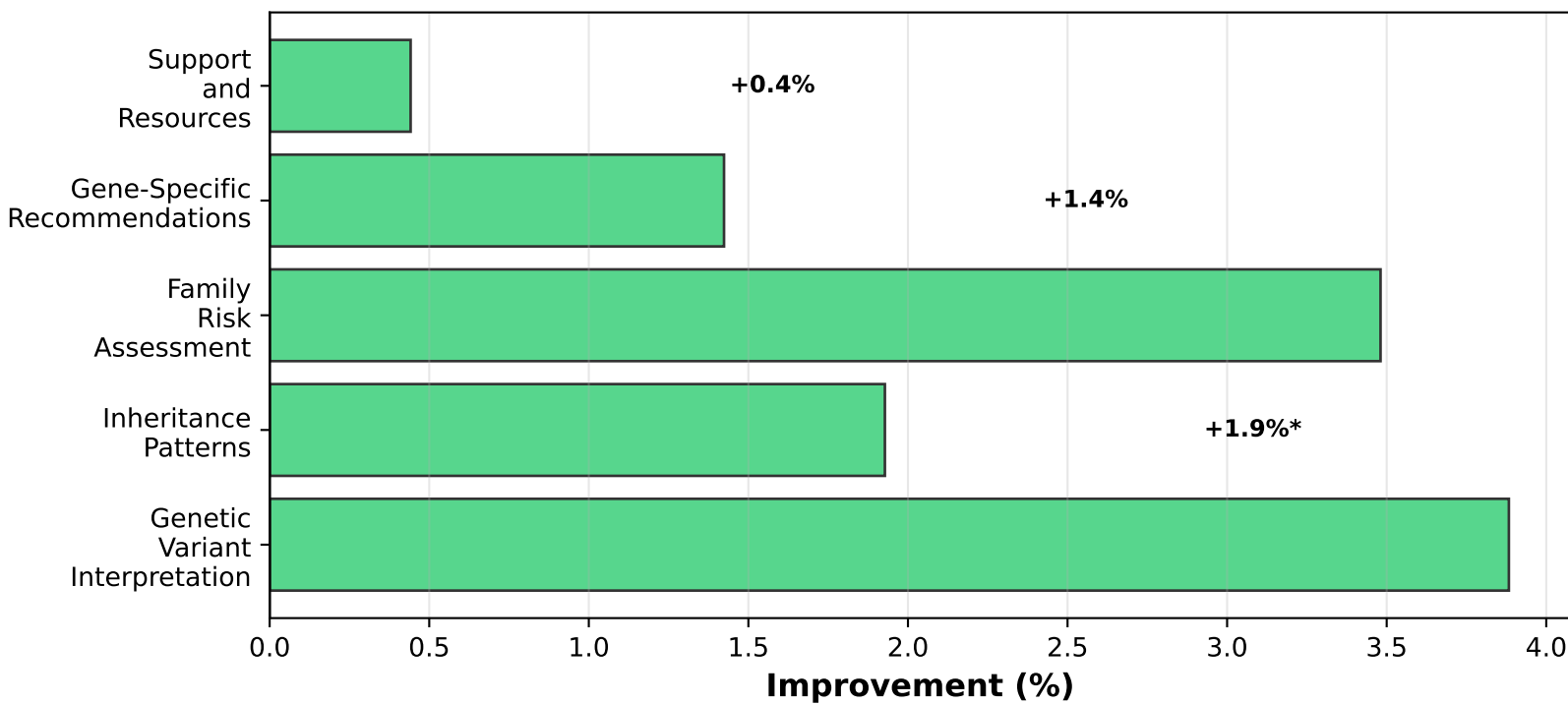
Overall Performance Comparison



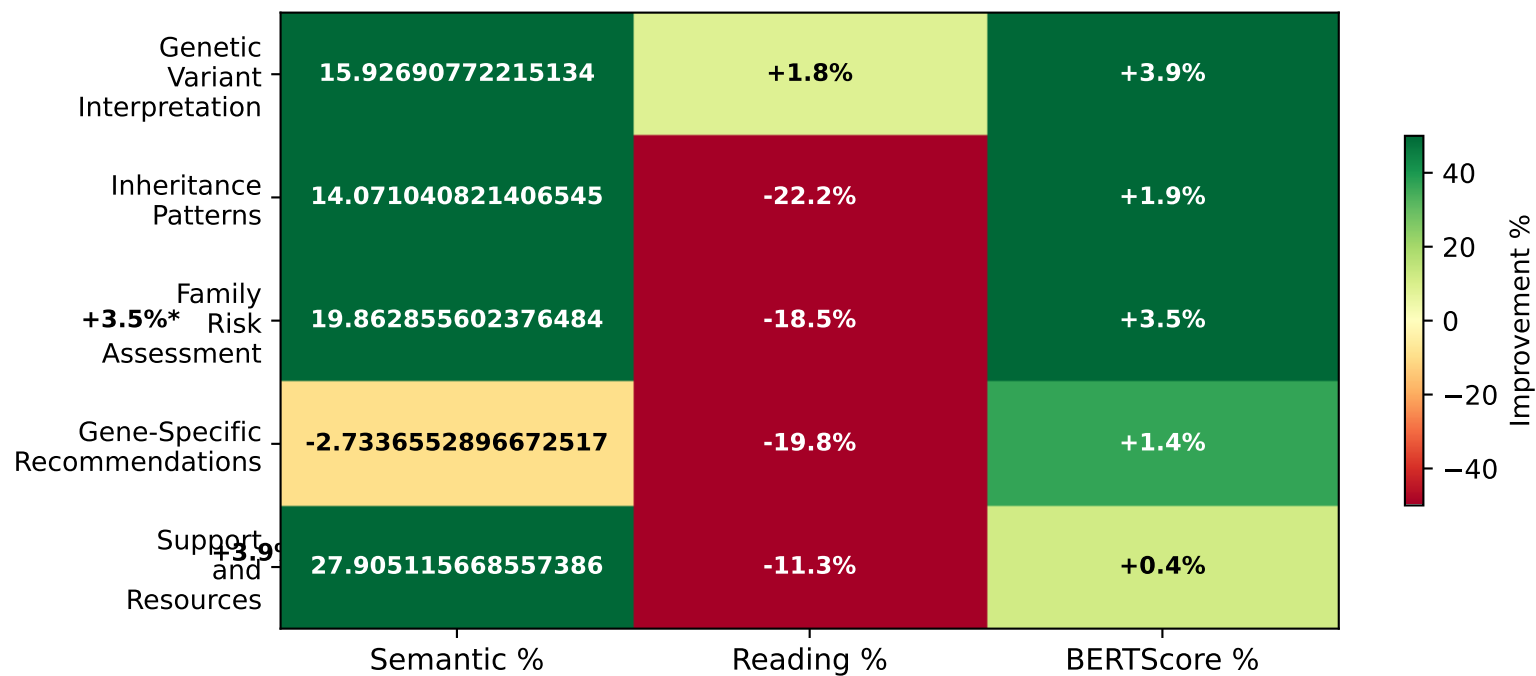
BERTScore F1 vs Gold Standard Distribution



BERTScore F1 Improvement by Category



Category Performance Matrix



Comprehensive Performance Analysis Summary

Metric	Raw Mean	RAG Mean	Change	P-value	Significant	Effect Size
Semantic Similarity	0.521	0.588	+12.8%	0.0043	✓	Medium
Reading Ease	35.9	30.2	-16.0%	0.0050	✓	Medium
BERTScore F1 vs Gold	0.748	0.763	+2.0%	0.0014	✓	Medium

Key Insights:

- Total: 49 questions across 5 categories
- Best performing category: Genetic Variant Interpretation (+3.9% BERTScore)
- Needs improvement: Support and Resources (+0.4% BERTScore)
- RAG shows consistent improvements in semantic alignment with gold standard
- BERTScore provides objective measurement against authoritative medical guidelines

Methodology:

- BERTScore calculated against 49 gold standard answers
- Statistical significance tested using paired t-tests
- Categories enable targeted analysis of genetic counseling topics
- RAG approach uses medical guidelines for enhanced accuracy